Supplementary Discussion

The MS^1 mass for the major MPC of rabbit pyruvate kinase (231,775 +/- 7 Da, **Supplementary Fig. 3a**, value of lowest-mass peak in the highest intensity cluster of peaks, *vide infra*), along with the MS^2 mass of the most abundant ejected subunit (57,941.4 +/- 0.5 Da, **Supplementary Fig. 3b**) and data from MS^3 subunit fragmentation (**Supplementary Fig. 3c**) were used as input into the new search engine (accessible online at http://complexsearch.kelleher.northwestern.edu). The result of this search was a homotetramer of N-terminally acetylated, -Met_{ini}, isoform M1 and a S401A with an extremely high MPC-score of 2,983 (**Supplementary Fig. 3d,** P11974, E-value of 2×10^{-51}). Careful examination of the MS^1 spectrum revealed the presence of multiple MPCs that differ in mass (**Supplementary Fig. 4a**). The ejected monomer from the highest abundance cluster of five MPCs (**Supplementary Fig. 3b, 4b**) was also present as a modified form observed at ~30% abundance (58,016.9 +/- 0.6 Da), +75.5 +/- 0.6 Da higher in mass. Isolation and fragmentation of this modified form (**Supplementary Fig. 4c**) allowed us to localize the modification as a β-mercaptoethanol covalently bound to Cys165 (+75.998 Da, theoretical Δm). Treatment of the sample with DTT for 1 hour removed all measurable amounts of this modification (data not shown), strongly indicating that it is in fact a cysteine modification and providing further confidence in its correct identification. Isolation and activation of the second, lower abundance cluster of MPC peaks ejected 4 unique monomers (**Supplementary Fig. 4b**), with the unmodified and modified forms previously identified along with a second, lower mass proteoform (55,729.8 +/- 0.8 Da) and a higher-mass modified form (55,805.0 +/- 1.0 Da) present at ~30% abundance. Isolation and fragmentation of the unmodified smaller form (fragment map shown in **Supplementary Fig. 4d**) allowed for its precise characterization as an N-terminally acetylated, truncated [23–531] form of pyruvate kinase. The N-terminal acetylation after truncation indicates that this endogenous cleavage occurred in the cell, and not during sample processing. The +76 Da modified form of the truncated ejected subunit was not sufficiently abundant to yield good fragmentation data, however its mass is consistent with the truncated form and the addition of the same mercaptoethanol modification present at ~30% occupancy. Together, these four subunit proteoforms formed a total of five tetrameric complexes differing in mass by ~76 Da (corresponding to between 0 and 4 mercaptoethanol modifications per tetramer); however, three of those species could contain a modification on the truncated form or the non-truncated form, making a total of eight possible MPCs. With the 5 MPCs from the higher-abundance cluster, we were able to characterize 13 total tetrameric MPCs present in the “protein standard” pyruvate kinase (**Supplementary Fig. 4a**),

highlighting the wealth of information that comes with MPC-level characterization of protein complexes.

Additionally, the previously unreported rabbit PGAM2 dimer complex (MS^1 of 57,156 +/- 3 Da for the most abundant set of MS^1 masses) was observed at low levels in the same PK sample with a confident E-value of 6×10^{-12} for the 20 matching fragments ions from its major ejected subunit (G1U7S4) and another high MPC-score of 2,983 (**Supplementary Fig. 5**). A satellite peak of the ejected monomer was observed at ~128 Da lower in mass and ~25% abundance relative to the major subunit proteoform. Fragment ions from this minor proteoform (co-isolated with the major monomer form and manually validated, data not shown) were consistent within 10 ppm with a C-terminal lysine cleavage (-128.18 Da). In total, the two forms of the ejected monomer formed three dimeric MPCs.

We also implemented the 3-tier tandem MS of MPCs to the human GAPDH complex fractionated from HeLa S3 cells (143,938 +/-1 Da), which was determined to be a homotetramer harbouring four P04406 subunits with an MPC-score of 2,924 (**Supplementary Fig. 6**). While only a single MPC was observable in the MS^1 spectrum, the MS^2 monomer ejection exhibited three forms that differed in mass by 31.0 +/- 0.9 Da. This mass shift is consistent with an endogenous cysteine persulfide modification (Cys-S-S-H, theoretical Δm: +31.972 Da), and the multiple monomer masses produced by a single tetrameric MPC are consistent with these modifications being partially labile to HCD fragmentation in the gas phase. Fragmentation of all forms present allowed for one of the modifications to be localized to Cys152, which is a known site annotated in UniProtKB as a persulfide; however, we were unable to localize the other modification precisely. We therefore considered only the mass of the singly modified monomer for use with the search tool; the 133.9 Da Δm is consistent within 6 Da of an additional four cysteine persulfide modifications ($4 \times 31.972 = 127.89$ Da). The extremely high MPC-scores determined for these homomeric complexes can be attributed to the lack of other possible homomeric complex stoichiometries; the next closest MPC is an entire subunit away (*e.g.*, ~58 kDa for PK) and is severely penalized in the current MPC-scoring framework.

Supplementary Tables

Supplementary Table 1. Various terms and their usage.

Term	Description	Example	Scoring	Reference
isoform	A specific base primary sequence, as the Uniprot Consortium uses the term; denoted with a "-1, -2, *etc.*" identifier after an accession number	P28074-2	E-value[a]	www.uniprot.org
proteoform	A specific molecular form of a protein produced from a specific gene	PFR2374	C-score[b]	Smith *et al.*, 2013.
multi-proteoform complex	A protein complex comprising specific proteoforms	Canonical 20S proteasome[c]	MPC-score	this work

[a] Q-values are also used.
[b] see LeDuc *et al*., 2014.
[c] 20S Proteasome with primary accession number (canonical sequence of the protein).

Supplementary Table 2. A description of parameters used to determine the MPC-score.

Mathematical Entity	Definition
MS^1 Likelihood	The value returned from the Generative Model of the MS^1 likelihood function. This is modeled as a truncated Gaussian centered at the theoretical mass, scaled to a maximum height of 1, with a standard deviation of 200 Da, and a minimum value of 1×10^{-300}.
Comment: This is the likelihood of observing the observed MS^1 value, if the complex in question was, in fact, in the instrument when the observation was made. Since it is a likelihood function, it takes an MS^1 mass value and returns a probability between zero and one. With this first implementation, a simple generative model was use where the theoretical mass of the complex represents a perfect match and has a probability of one. Values farther away from the theoretical mass are given increasingly lower probabilities, based on a Gaussian curve with standard deviation of 200 Da. This gives the model the essential property of considering complexes with MS^1 values greatly different from the observed theoretical mass, which is important given the huge number of potential unknown modifications. Lastly, a floor value is added to allow all complexes to be considered without rounding issues.	
Proteoform Likelihood	1 if the proteoform's C-Score > 50, 0 if the proteoform's C-Score < 0.5, otherwise the proteoform's posterior probability.

Comment: The MS² and MS³ data are used to identify the proteoform detected. The Proteoform Likelihood function splits the range of C-Scores into three logical categories; a proteoform with a C-Score over 50 is considered to have been identified and fully characterized, so the Proteoform Likelihood is set to 1; likewise, a C-Score below 0.5 is considered the completely wrong proteoform, and the C-Score is set to 0; for values between these two extremes, the Proteoform Likelihood function returns the Proteoform Posterior Probability generated in calculating the C-Score – this is a value between zero and one, with larger values representing increased confidence in the proteoform's identification.	
Complex Likelihood	$Complex\ Likelihood = MS_1\ Likelihood\ \times Proteoform\ Likelihood$
Comment: The two values above are multiplied together to give a likelihood for the whole complex based off the consideration of all three levels of MS data. One can easily imagine future iteration of this approach that use more complex relationships between these two values, to increase the sensitivity of the overall system.	
Marginal Likelihood	$Marginal\ Likelihood = \sum_{i=1}^{n} Complex\ Likelihood_i$
Comment: The sum of all Complex Likelihoods interrogated.	
C-score	Characterization score on a Phred-like scale.
Comment: The C-score is defined in LeDuc et al. 2014, and is used as a measure of the uniqueness of the characterization of a proteoform. Any score above 50 is taken as a highly confident characterization.	
Posterior Probability	The proteoform's posterior probability, calculated during the process of the above C-Score.
Comment: The C-score is calculated within a Bayesian framework, so the proteoform's posterior probability is the probability that the proteoform was in the mass spectrometer at the time the measurements were taken, after considering the observed MS data. This is as opposed to the prior probability which is the probability that the proteoform in question was in the mass spectrometer before considering the MS data.	
Complex's Posterior Probability	$Complex\ Posterior\ Probability_i = \frac{Complex\ Likelihood_i}{\sum_{j=1}^{n} Complex\ Likelihood_j}$
Comment: A given complex's posterior probability is simply that complex's likelihood divided by the sum total of the likelihoods of all complexes considered.	
MPC-score	$MPC\ score = -10 \times log_{10}(1 - Complex\ Posterior\ Probability)$
Comment: This function rescales each complexes' posterior probability to a more convenient value. Effectively the MPC-score is ten times the number of nines following the decimal point in the complex's posterior probability. For example CPP = 0.9 will yield an MPC-score of 10, while 0.99 is 20, 0.999 is 30 etc.	

Supplementary Table 3. Results from denaturing top-down of subunits obtained via automated LC-MS/MS of the human 20S proteasome.

Name	Intact mass (Mono.)	Intact mass (Avg.)	p-score	C-score	# of matched fragments	Accession #	Isoform	Modifications	Comments
Alpha 1	Not Observed	Not Observed	N/A	N/A	N/A	P25786	1	N-T[a] Acetyl	Isoform determined from cleaved product (Biomarker hit)
Alpha 2	25,793.24	25,809.44	7.1×10^{-33}	461.0	25	P25787	1	-Met_{ini}, N-T Acetyl	
Alpha 3	28,406.02	28,424.06	6.1×10^{-6}	20.1	6	P25788	1	-Met_{ini}, N-T Acetyl, monophosphorylated	
Alpha 4	29,376.15	29,394.66	2.4×10^{-40}	477.6	23	P25789	1	-Met_{ini}, N-T Acetyl	
Alpha 5	26,436.22	26,453.07	2.0×10^{-29}	778.1	18	P28066	1	N-T Acetyl	
Alpha 6	27,292.78	27,310.30	2.5×10^{-71}	822.5	49	P60900	1	-Met_{ini}, N-T Acetyl	Xtracted from a single scan and searched with Prosight Lite
Alpha 7	27,780.58	27,797.70	2.6×10^{-46}	442.8	48	O14818	1	-Met_{ini}, N-T Acetyl	
Beta 1	23,533.94	23,548.97	4.1×10^{-47}	594.9	35	P20618	1		
Beta 2	22,863.69	22,878.32	9.6×10^{-29}	1,030.8	17	P49721	1	N-T Acetyl	
Beta 3	22,826.46	22,841.70	4.2×10^{-32}	3.0	27	P49720	1	-Met_{ini}, N-T Acetyl	
Beta 4	24,376.12	24,391.78	9.0×10^{-20}	343.9	20	P28070	1		
Beta 5	22,444.12	22,458.37	1.9×10^{-46}	621.8	32	P28074	1		
Beta 6	21,889.80	21,903.89	2.2×10^{-5}	15.9	8	P28072	1		Major form oxidized
Beta 7	25,278.79	25,294.99	2.3×10^{-12}	234.3	12	Q99436	1		

[a] N-terminal

Supplementary Table 4. The sets of isoforms and their copy numbers determined for MPCs of the human proteasome resulting from four different searches using MS^3 data obtained from four different proteasome subunits (alpha-4 and alpha-5 were co-isolated). Better scoring results from other MPCs (PA28–20S proteasome and PA28gamma–20S proteasome) were not considered. Best hit MPCs all had poor MPC-scores <<1. Results are displayed for each of 14 subunits as [Isoform #, Copy #].

	Searched Subunit			
	Alpha 4	**Alpha 5**	**Alpha 7**	**Beta 6**
Delta mass (Da)	35	36	35	3
MPC-score	3×10^{-4}	3×10^{-4}	4×10^{-4}	3×10^{-4}
a1	1,2	1,2	2,2	2,2
a2	1,2	1,2	1,2	1,2
a3	1,2	1,2	1,2	2,2
a4	1,2	2,2	1,2	1,2
a5	1,2	1,2	1,2	1,2
a6	1,2	1,2	1,2	1,2
a7	1,2	1,2	1,2	1,1
b1	1,2	1,2	1,2	1,2
b2	1,2	1,2	1,2	1,2
b3	1,1	1,2	1,1	1,2
b4	1,2	1,2	1,2	1,2
b5	2,2	2,1	2,2	1,2
b6	1,2	1,2	1,2	1,2
b7	1,2	1,2	1,2	1,2

Supplementary Table 5. Neutral monoisotopic masses for use with the web tool.

See attached file: Supp_Table_5_NeutralMasses.xlsx